图书情报工作
第 65 卷 第 7 期 2021 年 4 月

世界中国学专题数据库建设与应用研究*
——来自华东师范大学的实践

■ 邓璐芗[1] 许鑫[1,2]
[1] 华东师范大学调查与数据中心 上海 200241 [2]华东师范大学经济与管理学部 上海 200062

摘 要：[目的/意义]通过建设专题数据库解决当代世界中国学研究存在的域外中国知识难以获取、对舆情重视程度不够、忽视人物间关联等问题。[方法/过程]从数据采集、加工整合和检索利用阶段，分别论述世界中国学专题数据库的设计方案，并以华东师范大学世界中国学数据库为例，介绍其数据采集、加工整合和检索利用情况。[结果/结论]通过建设世界中国学数据库，可以促进对在域外进行中国问题研究的重要人物、重要机构、重要研究领域以及人物间关联的发现。此外，该数据库还为学者们开展更深层次、更多维度的相关研究提供数据基础。

关键词：世界中国学 专题数据库 智库 舆情
分类号：G250
DOI：10.13266/j.issn.0252－3116.2021.07.005

近年来，中国的快速发展取得了举世瞩目的成绩，国际社会对中国的关注度持续升高。随之而来的全球范围内关于中国的知识生产也迅速激增，国外媒体、学者及机构对中国问题的研究越来越多，大到宏观层面的社会、经济、政治，细到特定年代的吃穿住行，研究内容涉及中国的方方面面。这些研究的目的在于获取有关中国的重要知识和情报，以便在与中国的合作与竞争中取得优势。由于信息不对称、意识形态、政治对抗等多方因素，导致这些研究成果往往带有偏见，对中国的国际形象有所损害，也不利于营造友好的国际发展环境。这一问题在新冠疫情期间表现得尤为突出：中国在抗击疫情方面取得了重大成就，可为其他各国的疫情阻击提供可资借鉴的宝贵经验和重要支持，但西方社会却出现了刺耳的声音，部分政客和媒体罔顾事实，一味抹黑中国及中国抗疫经验，指望发起新一轮"冷战"遏制中国发展。在这样的时代背景下，系统地分析和研究全球范围内关于中国的知识，知已知彼，方能在国际合作与对抗中不受制于人，维护己方利益。

外国学者对中国的研究，主要包含了两大主题：①对中国历史、文化和语言的研究，通常称"汉学"（Sinology）；②对现当代中国问题的研究，涉及中国的政治、经济、军事、社会等方方面面，称"中国学"（China Studies）[1]。"汉学"研究曾是域外中国研究的主流，这一境况在第二次世界大战后被颠覆。随着新中国的成立和崛起，对现当代中国问题的研究越来越受到各国研究者的重视，由学术问题进而演变为政治、外交、军事和经济等领域的重要议题，研究成果也突破了学术范畴，对其所在国决策影响显著。外国学者对中国的研究，为国内学术界开展相关问题的研究提供了新思路、新方法、新成果。但不可否认的是，在这些研究成果中，存在着许多带着疑问的、牵强附会的，甚或是错误的、别有用心、歪曲事实的研究。对于新思路、新方法和新成果，有必要取其所长，解决国内的问题；对于疑问，需要及时的回应，答疑解惑；对于攻击，需要精准有利的回击。这需要国内的学者，重视域外学者对中国问题的研究，去了解其研究的动态进展和观点。

1 世界中国学相关研究概述

针对域外对中国问题的研究，学术界有海外中国学、国外中国学、域外中国学、国际中国学、世界中国学

* 本文系华东师范大学"幸福之花"先导研究基金项目"大数据视阈下基于学术共同体的人文社科学术评价与促进研究"（项目编号：2019ECNU－XFZH016）研究成果之一。
作者简介：邓璐芗（ORCID：0000－0002－8703－7468），助理馆员，硕士；许鑫（ORCID：0000－0001－7020－3135），调查与数据中心副主任，教授，博士，博士生导师，通讯作者，E-mail：xxu@infor.ecnu.edu.cn。
收稿日期：2020－10－26 修回日期：2021－02－05 本文起止页码：50－59 本文责任编辑：徐健

等多种称谓。朱政惠提出,海外中国学指外国学者对中国的方方面面的研究,除了对中国历史的研究,还有对中国文学、哲学的研究,以及对中国政治、经济、文化、社会、军事等多方面的探讨[1]。梁怡认为国外中国学研究指国外有关中国近代、现代、当代哲学社会科学领域的研究[2]。域外中国学、国际中国学的概念与前两者相近。近年来,上海社会科学院提出"世界中国学"的概念,认为世界中国学指海内外以中国为对象的包罗万象的综合性研究[3],在研究内容上涵盖了前述概念下的"汉学"和"中国学",还包含了立足本土的中国问题研究。世界中国学逐渐发展成为一门新兴学科,其目的在于促进域外和本土研究的互动,真正助力构建中国特色话语体系和知识体系。而与之相对的世界中国学研究,则指的是国内对全球范围内中国问题研究的反研究。

在传统的"汉学"领域,朱政惠介绍了柯文(P. Cohen)教授中国史学研究的"中国中心观"形成过程,并且指出,柯文教授认为与海外中国研究队伍的规模相比,中国学者的反研究显得不够对称[4];严绍璗对日本中国学的诸多研究者和其著作进行分析解读,认为中国研究者要在外国研究者对中国进行研究的文化语境、文学史观念和文本的原典性观念方面做出反思,并调整学术视角[5];吴原元基于前人的研究成果,将中国学人对域外汉学的态度分为清末明初基于学术自信的轻蔑和褒奖、1920 年代后的取法和纠谬、1949 年后对其所隐含的意识形态和方法论等进行批判、1980 年代后的对西方的迷信和盲从 4 个阶段[6]。可以看出,对传统域外"汉学"的研究,国内学者主要是围绕某个具体的研究者以及他们的研究著作开展,采用的是精读、深度解析等传统的人文研究手段。

随着在其他国家以当代中国问题为主要目标的中国问题研究愈来愈得到重视,国内学者也逐渐意识到对域外相关研究进行反研究的重要性。朱政惠指出,费正清(J. Fairbank)开创的美国现代中国学,不仅是美国的中国研究的重要转向标识,也是世界中国问题研究转向的分水岭[1];魏海生、唐洲雁、张树华认为,讲好中国故事、传播好中国声音,可以影响在域外进行中国研究的研究者,进而影响整个世界正确认知中国[7];梁怡在研究中指出,在域外的中国问题研究中,智库对政府的影响力逐渐代替了高校研究机构的作用[2]。

对当代中国问题的重视,导致世界中国学的研究内容和研究主体发生了巨大的变化,内容上由以历史和文学为主,转为了以政治、军事、外交、社会、经济等问题为主;研究主体上由以学者为主,转为了政、军、商、媒、民共同参与,甚至研究主体所处的国别,也由英法美日等发达国家和邻国,扩张到了更大的范围。

2 世界中国学研究面临的困境

相对全球海量的中国问题研究和知识生产,国内学者在进行反研究时仍面临着如下困境:

2.1 对研究的重要性和价值认识有限

目前,已有近百个国家设置有相当规模的中国问题研究机构,遍及世界各地,这些机构通过多种途径掌握中国动态,服务于学术研究和国家决策。而与之相对,国内学者对域外中国问题研究的反研究,仍是介绍多于分析,翻译多于研究[8]。这使得部分学者在不知不觉中沦为了西方学术和观点的传声器,失却了学术的自主性和独立性。这与我们倡导构建富有中国特色的话语体系的愿景相违背。究其缘由,是没有真正认识到跟踪、研判域外中国研究动态的重大价值。在学术方面,世界中国学研究可以借鉴国外优秀的研究方法和研究成果,促进我国相关问题的研究,亦可发现疑问和错误,有针对性地释疑解惑;在服务国家决策方面,世界中国学研究则是以国家为行为主体,是掌握国际重大议题设置的主动性、构建中国特色话语体系、寻求共识共谋发展的前提条件,具有重要意义。

2.2 难以获取全面、及时的域外中国知识

国内的世界中国学研究大多围绕某个人物、某篇著作、某一言论或新闻报道进行文本解读。在互联网时代信息飞速传播的大背景下,一是难以及时获取到最新的信息,使得研究内容的时效性得不到保障,域外关于中国的疑问、关于中国的错误观点和责难必然得不到及时、有力的回应,势必对中国造成不良的影响;二是难以获取全面的域外中国知识,无法构建系统、全面的域外认知全景,使得研究结论存在以偏概全的可能性,无法保证学术的客观性以及对决策的有效支持。

2.3 对舆情重视程度不够

智库承担着对内提供政策咨询和建议、引导媒体和公众,对外扩大国家影响力和话语权的重要职能,其反映出的舆情对各国的政府治理、社会经济、军事外交都有着重大的影响力。同时,新闻媒体和社交网络所承载的舆情具有传播广泛、影响面广等特点。随着全球化进程的加快,在国际政治对抗和交流中,国外舆情中议论及传播的对中国问题的看法和观点,对下至普通的外国民众,上至政府决策层的影响都愈加显著。需要特别注意的是,国内的世界中国学研究仍主要强

调学术研究的重要意义,服务于现实需要的研究较少,尤其是对国外智库成果的研究程度,难以匹配当今智库的重大影响力。

2.4 忽视对域外人物间关联的研究

在域外进行中国问题研究的学者中,有部分学者的影响力卓著,如费正清、傅高义(E. Vogel)、史华兹(B. Schwartz)、柯文、村山吉廣、吉川幸次郎等。这些学者对所在国针对中国的学术研究和政治决策都有着很大的影响力,并且许多学者之间关系密切,如柯文就曾在哈佛大学费正清东亚研究中心任研究员,还深受史华兹的影响。我们进行世界中国学研究的初衷,除了借鉴观点和补充不足之外,还包括有针对性地就域外中国研究所关注的问题展开对话交流,引导其正确认识中国。这就需要我们不仅识别出域外中国问题研究的关键人物,还要找出将中国声音传递给这些关键人物的路径。最终通过这些关键人物影响其所在国的决策和民众的认识。

在国际形势风云变幻、反华浪潮一再掀起的时代大背景下,以上所列严峻问题亟需有效解决,才不至于在国际合作与竞争中因相关信息掌握不利而处于劣势地位。

3 世界中国学专题数据库设计方案

3.1 建设意义

大数据时代的到来极大地丰富了世界中国学研究所能倚靠的信息资源。但同时,大数据又具有规模大且生产迅速、类型多样、质量参差不齐且整体价值密度低等特点,使得传统的信息方法和技术无法应对海量异构数据的检索、加工和分析的需求。由此,基于智能技术对海量数据进行处理、分析和挖掘的"数据智能"应运而生。需要说明的是,"数据智能"方法和技术的学习及掌握难度较高,因此,许多世界中国学研究者面对大数据时难免会望而生畏。在这样的背景下,就有了专题数据库的用武之地。

专题数据库是以某一学科、领域、人物、事件等为对象而建构起来的专门性数据资源集成[9],是人文社科数据资源开发与利用的创新手段和重要途径,在学术资料提供、战略决策支持、传统文化传承等方面发挥着重要作用[10]。其面向特定学科、领域、人物或事件等对象的特点及研究人员的特色需求,同时也具有信息集中、内容专深、价值密度大等特点。世界中国学专题数据库主要对全球范围内针对中国研究的各类研究成果进行收集、分析、处理、存储和展示,其建设有助于学者更为便利地获取该主题内全面的数据资源开展研究,既是对该主题领域研究重要性和价值的认可,也是促进相关主题研究良性发展的重要举措。一方面,学者们可以基于专题数据库中的相关资源直接开展研究,消除了学者获取并整合海量多源异构数据资源的痛苦;围绕对舆情重视程度不够的问题,将智库、新闻媒体及社交网络的相关信息资源整合入世界中国学专题数据库中,正好可填补这一空缺;针对忽视对域外人物间关联的研究,在世界中国学专题数据库的建设中,可针对性地以人物为核心,发现并可视化展示人物间关联,有助于依据人物之间的关联规划中国声音的传递路径,支持对国家决策至关重要的舆情进行精准出击或有力回应;此外,对域外中国研究相关知识的全方位展开,还有助于不同学科领域的学者加入世界中国学研究这一领域,为国内学界带来世界中国学研究新的解读与利用方式。

3.2 建设思路与难点问题

人文社科专题数据库的构建通常要经过资源采集、加工整合及检索利用多个阶段[11]。世界中国学数据库的建设,需要在这些必经阶段中解决当前世界中国学研究面临的痛点难点问题。

3.2.1 数据采集

数据资源是数据库的核心,科学合理地选择数据来源并确定采集策略,决定了数据资源是否能够有效地支撑世界中国学领域的各项反研究。对当代中国问题的重视,要求世界中国学专题数据库的数据资源不仅包含学术研究成果,还要重点关注智库、新闻媒体、社交网络等来源的信息资源。学术研究成果通常分布在各类学术数据库中,具有数据字段丰富、规范等特点。智库、新闻媒体等来源的信息,分布广泛、异构情况严重。社交网络信息通常集聚在少数大型社交网络平台之上,信息具有短小等特点。单一化的数据采集策略难以兼顾这些来源广泛、差异性大的信息资源,使得数据采集成为世界中国学专题数据库建设面临的重点和难点问题之一。

因此,需要针对不同类别信息的特征,分别制定不同的数据采集策略。针对学术研究成果,可先行选取有效检索词在学术数据库进行检索,再基于检索结果使用爬虫工具进行数据爬取;针对智库和新闻媒体信息,可采用通用引擎与主题爬虫相结合的采集方式,首先选取适当的关键词形成主题词表,利用通用搜索引擎广度优先的抓取策略进行互联网特定信息采集,再使用垂直搜索主题爬虫对重要程度高的权威网站进行

深度抓取，实现对信息采集范围和方向的有效控制；针对社交网络信息，其数据采集可采用深度优先搜索方法，先采集帖子或话题的标题、链接、作者等信息，继而进入帖子或话题内部采集回复信息、链接、作者等信息，再进入回复的内部采集楼中楼的回复、作者等信息，如此遍历所有能采集到的相关数据。

3.2.2 数据加工整合

世界中国学专题数据库的数据加工整合，包括数据清洗和数据聚合。

数据清洗的要点主要包括去重、乱码字段及格式清洗、别名统一等。世界中国学研究的目的和初衷之一，在于就域外中国研究的重点关注问题展开对话交流。而人作为对话交流的主体就显得尤为重要。在数据清洗过程中，尤其重视对一人多名情况的处理。英文姓名组织形式多样，如存在全名、去掉中间名、仅中间名缩写、除姓之外均缩写、姓与名之间有间隔（采用圆点或横线间隔）等情况。人物姓名数据如何清洗和呈现是数据清洗的难点。因此，在数据加工整合时，需要统一人物名称，将同一人物的多个别名进行映射。

当前，数据资源聚合总体呈现出由基于外部特征到基于内容的聚合、由粗粒度到细粒度的聚合、由单一维度到多个维度聚合的转变趋势[12]，可供应用的技术和方法较多，常见的如主题词表、本体、关联数据、文献计量、分众分类、社会网络分析[13]等方法，也有学者提出基于知识超网络理论的数据资源语义聚合模型[12]等新方法。这些方法各有优劣，需要依据资源的具体情况进行选择。此外，世界中国学研究关注人物、问题及观点3个重要要素。如何将类型多样、结构差异大的各类信息整合到数据库中并有效呈现出人物、问题及观点相关信息，是世界中国学专题数据库建设的关键。考虑到人作为提出问题和发表观点主体的重要性，在数据整合时应以人物为核心，将人物与人物提出问题或发表观点的相关动作（如学术论文和智库文章的发表、社交网络发言及回复）进行关联、组织及呈现，以便后续发现问题和精炼观点的研究推进。

3.2.3 数据检索利用

在检索方面，检索效果主要体现在准确性和效率两方面。传统以关键词匹配方式进行的检索，其检索过程中易出现"忠实表达""表达差异"和"词汇孤岛"等问题[14]，从而影响检索结果的准确性。此外，检索过程中如缺乏推理，会增加用户通过检索获取知识的难度，使得检索的效率低下。为此，不少学者提出了语义检索的概念，如借助实体识别、实体分类、关系抽取、知识存储、知识推理等语义化技术实现灵活的语义检索[11]，可以有效帮助用户快速获取所需资源。

在利用方面，可对相对浅层次的信息进行分析并通过直观的可视化界面予以呈现，如关键人物、重要机构、人物之间关联、人物与问题的映射关系、人物与观点的映射关系等，供相关的学者及决策者直观的查阅，他们就可将有限的时间和精力投入到更复杂和纵深问题的研究上，从而减轻其负担。目前，可供利用的可视化方法和技术较多，如关联网络、多层网络、知识图谱等，可结合具体数据情况及用户需求进行选择。

3.3 平台架构

建设世界中国学专题数据库需要解决如下重点问题：①采集、挖掘域外中国问题研究的人物及其问题、观点，并且多维度呈现域外中国研究的全貌；②基于门户网站，提供多样化检索功能，便于查找人物、成果及观点；③通过可视化技术直观地展现人物之间的复杂关联，帮助学者和决策层更为便利地制定信息传递路径和机制。其架构自下而上主要包括基础设施、技术支撑、数据库、数据应用和服务展示，如图1所示：

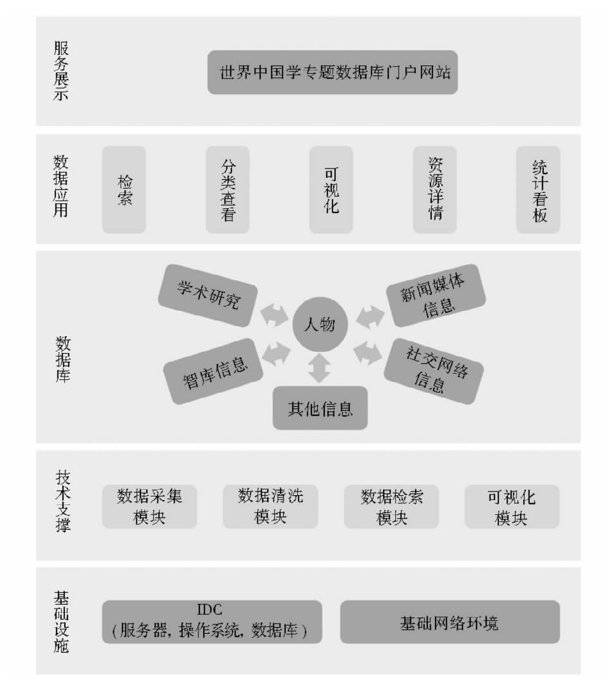


图1 世界中国学专题数据库平台架构

（1）基础设施。世界中国学专题数据库的基础设施包含服务器、数据库、基础网络等内容。数据库可选择CentOS作为操作系统，安全、低维护、稳定；数据库类型可选用PostgreSQL关系数据库，开源、功能丰富。

（2）技术支撑。世界中国学专题数据库的技术支撑包含数据采集、数据清洗、数据检索、可视化等模块，使用

的技术包含 Python 爬虫、Java Web 服务、前端可视化等。

(3)数据库。在世界中国学专题数据库中,可汇聚学术研究、智库、新闻媒体、社交网络等多来源、多类型的信息,以人物为核心,建立人物与各类信息之间的关联,为更深层次的问题及观点提炼、以及信息传播路径的规划奠定基础。

(4)数据应用。世界中国学专题数据库的数据应用包含数据查看、检索、可视化等功能。这些功能将数据按照图形或文字方式展示给用户,并且提供选择、检索的方式进行更好的展示。

(5)服务展示。通过世界中国学专题数据库门户网站进行服务展示,汇聚所有的数据应用功能。

4 华东师范大学世界中国学数据库平台建设及应用

华东师范大学作为国内最早对域外中国问题研究进行反研究的科研机构之一,于 2019 年建设世界中国学专题数据库,以支持政策决策,推动世界中国学研究的进一步发展。该数据库是对中国区域外进行中国研究的人物及其相关论文、著作、报告、新闻、智库成果等各类分散资源进行收集、分析、处理、存储和展示的信息资源数据库。

4.1 数据采集

4.1.1 数据来源

为同时覆盖“汉学”和“中国学”两大主题领域的域外中国研究成果,专题数据库选取谷歌学术和智库两大数据来源,其中谷歌学术兼顾了“汉学”和“中国学”,而智库则主要针对的是政治、军事、社会、经济、科技等当代中国问题。

4.1.2 数据采集策略

有别于传统数据采集中直接采集通过关键词检索的论文、报告等数据资源,在采集该数据库所需数据资源时,考虑到降低学者库中学者去重、清洗的难度和压力,选择从检索列表中析出学者信息,再通过学者对其中国问题研究的相关论文、报告、新闻等资源进行扩展和搜集。

首先,在谷歌学术上使用 Chinese military、Chinese investment、Chinese conflict、Chinese language、Chinese internet、Chinese censorship、Chinese population、Chinese government、Chinese traffic、Chinese agriculture、Chinese industry 等检索词进行检索,获得军事、投资、语言、互联网等领域的中国研究相关学术成果;其次,从上述相关学术成果列表中,析出学者信息,获取学者字段;最后,获取学者所有论文列表及被引次数等数据,以 China、Chinese 为检索词从列表中筛选出与中国议题相关的论文成果,作为论文库数据来源。

针对智库,采用如下步骤获取数据:

(1)参照宾夕法尼亚大学于 2019 年 1 月发布的《2018 年世界智库影响力排名》(2018 Global Go To Think Tank Index Report)[15],在排名前 100 的世界智库列表中,从第一名开始使用 China 和 Chinese 作为检索词,按顺位选取有检索结果的前 20 个智库作为数据来源,分别如表 1 所示:

表 1 智库数据来源

全球排名	智库	全球排名	智库
1	(美)Brookings	21	(德)SWP-Berlin
9	(美)RAND Corporation	2	(法)French Institute of International Relations, IFRI
7	(英)Chatham House	17	(美)Council on Foreign Relations
33	(韩)Korea Institute for International Economic Policy, LIEP	8	(美)The Heritage Foundation
3	(美)Carnegie Endowment for International Peace	12	(美)Peterson Institute for International Economics
11	(美)Wilson Center	31	(加)Center for International Governance Innovation, CIGI
5	(美)Center for Strategic and International Studies, CSIS	50	(德)German Council on Foreign Relations
47	(意)Italian Institute for International Political Studies	29	(美)Human Rights Watch
10	(英)International Institute for Strategic Studies, IISS	28	(英)European Council on Foreign Relations
64	(澳)Lowy Institute for International Policy	56	(澳)Australian Institute of International Affairs, AIIA

(2)分别进入以上智库官网,使用 China 和 Chinese 作为检索词,搜索相关学者。2019 年 10 月,以上 20 个智库中共有 3 168 位学者,以 China 和 Chinese 作为检索词搜索出的学者 240 名,占比约 7.58%。

(3)进入学者主页,获取学者字段,使用 China 和 Chinese 为检索词从学者的报告、观点、新闻、博客等资源中筛选出与中国相关的部分,作为舆情库数据来源。

4.2 数据加工整合

根据数据的实际情况,进行论文去重、发表日期统一、人物姓名统一、空字符处理、乱码字段纠正及格式清洗、重要计数清洗统计排序等。通过观察发现,学者个人详情页面上通常出现的是全名(full name),而其

他各处则经常出现去掉中间名(middle name)或缩写的情况。因此，在数据清洗时，将人物详情页面上的姓名，与其论文、智库文章中的署名进行匹配并添加映射关系。而针对部分学者出现多字段空值或无论文、智库成果的情况，考虑到此类学者在中国问题研究领域大概率处于边缘位置，故直接去除此类学者。

经清洗和整理，共获得学者、论文、舆情信息条目数量如表2所示：

表2　清洗后的学者、论文、舆情条目数量

单位/条

学者信息	论文信息	舆情信息
5 321	165 625	4 603

在数据库的建设过程中，对谷歌学术和智库两类数据来源进行预研究，并结合采集到的论文和舆情数据资源的字段情况，设计学者库表、论文库表和舆情库表，详细内容如表3、表4和表5所示，并分别构建学者库、论文库和舆情库3个子库。为建立人物与信息之间的关联，将论文库表与学者库表、舆情库表与学者库表通过字段scholarid进行关联。在对已清洗数据进行入库操作时，根据数据库表间的关联关系，查找人物的id(字段scholarid)，在插入论文和舆情数据时，使用对应人物的id作为填充字段，从而实现了3个子库间的联通。

表3　学者库表

序号	字段	说明	序号	字段	说明
1	id	主键	11	refernumber	引用次数
2	createtime	创建时间	12	refernumberyear	引用年份
3	imageurl	头像地址	13	key	唯一标识
4	name	姓名	14	source	来源
5	position	职位与机构	15	gscrsbst	h指数、i10指数
6	introduction	简介	16	initials	首字母，排序使用
7	updatetime	更新时间	17	summary	介绍
8	status	状态	18	refernumberocsd	中国问题研究引用总数
9	validatetype	验证方式	19	maprelationship	作者关系图谱
10	sphere	领域			

表4　论文库字段

序号	字段	说明	序号	字段	说明
1	id	主键	15	publisher	出版商
2	academicartices	学术搜索的文章	16	refernumberurl	引用数url
3	articleurl	原文链接	17	reportnumber	报告编号
4	author	作者	18	rollnumber	卷号
5	carwurl	采集的URL	19	seminarpaper	研讨会论文
6	createtime	创建时间	20	source	来源
7	fileurl	原文地址	21	status	1:正常显示
8	initials	首字母	22	summary	简介
9	institutions	机构	23	title	标题
10	journal	期刊	24	updatetime	更新时间
11	library	图书	25	refernumberclass	引用数分类
12	pagerange	页码范围	26	scholarid	学者id
13	periodnumber	期号	27	refernumberyear	引用数年份
14	publishtime	出版时间	28	publishtimeclass	发表年份分类

表5　舆情库字段

序号	字段	说明	序号	字段	说明
1	id	主键	8	source	来源
2	author	作者	9	status	正常状态
3	contents	内容	10	subject	主题
4	createtime	创建时间	11	title	文章标题
5	link	原文链接	12	type	类别
6	publishtime	发表时间	13	updatetime	更新时间
7	scholarid	学者id			

4.3 数据检索利用

4.3.1 查询检索

在世界中国学专题数据库中，既可在数据库首页设置对应的数据库并进行关键词检索，也可进入学者库、论文库或舆情库后再进行关键词检索或分类查阅。

在学者库中，支持按照学者来源及姓名首字母排序进行查阅。学者详情包括姓名、职位与机构、研究领域、来源、与其他学者的关系图谱、被引次数可视化、中国问题研究相关的论文或舆情列表。点击论文名称即可跳转至论文库中相应论文详情页面，查看论文的详细信息；点击舆情文章的名称即可跳转至舆情库中相应舆情详情页面，查看舆情的详细信息。

在论文库中，支持按照发表时间和被引次数进行查阅。论文详情包括标题、作者、简介、期刊、发表时间及卷期、页码范围、出版商、被引次数可视化等内容。

在舆情库中，支持按照舆情的类别进行查阅。舆情详情包括标题、作者、类别、发表日期以及舆情全文内容、原文链接等内容。

4.3.2 重要学者发现

在世界中国学专题数据库中，可视化地展示了被引次数 TOP10 的学者，分别为 M. F. Zhou、M. Santosh、B. M. Jahn、B. Naughton、J. G. Liou、M. H. Bond、B. Popkin、F. L. Wu、J. He、D. Shambaugh。因智库来源的学者没有被引次数的相关信息，故次数 TOP10 学者全部为论文学者。这些学者的发文数及被引次数历年变化情况如图 2 所示：

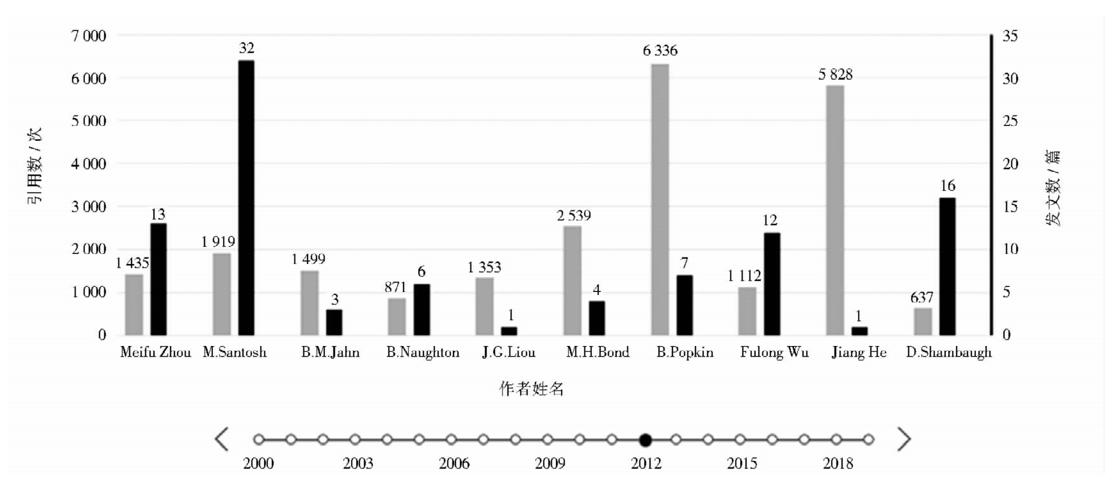


图 2 被引次数 TOP10 学者发文数及被引次数历年变化情况（示例）

经分析还可发现，这些学者的研究领域主要集中在地质学（4 人）、经济学（1 人）、社会心理学（1 人）、健康与营养学（1 人）、城市发展与治理研究（1 人）、流行病学（1 人）、亚洲国际关系研究（1 人）等领域，领域跨度大，而且地质领域不仅学者较多，相关研究产出也多，被引次数非常高。

4.3.3 学者关系图谱可视化

在世界中国学专题数据库中，可视化地展示了学者之间的关系图谱。图 3 以来自哈佛大学的学者 A. L. Johnston 为例，展示他本人与其他学者的关联，以及与他发生关联关系的学者之间的关联，如 A. L. Johnston 与 A. Carlson、J. J. Suh、P. Katzenstein 等人有关联，并且 A. Carlson 与 J. J. Suh、P. Katzenstein 3 人间互有关联，这说明 A. L. Johnston、J. J. Suh、A. Carlson、P. Katzenstein 4 人可能存在共同进行研究的情况。

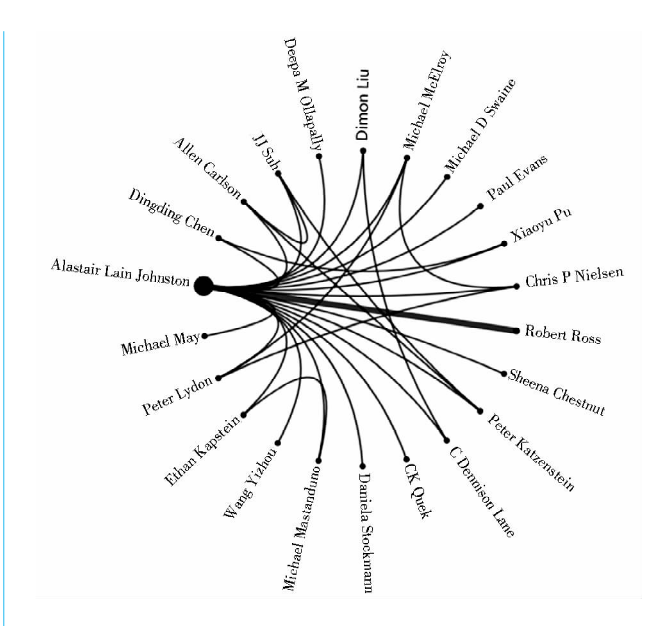


图 3 Alastair Lain Johnston 的学者关系图谱

4.3.4 重要研究领域发现

在世界中国学专题数据库中，可视化地呈现了世界中国学相关学者的研究领域，如图 4 所示。通过对学者及其成果所涉领域进行分析发现，Economics（经济）、In-

ternational Relations（国际关系）、International Business/Trade（国际贸易）、Political Economy（政治经济）、Machine Learning（机器学习）、Climate Change（气候变化）、Epidemiology（流行病学）等领域是域外学者关注的重点。值得注意的是，被引次数 TOP10 学者中有 4 人从事的是 Geology（地质学）研究，从域外中国研究整体情况来看，则仅是其中的一小部分。分析其原因，可能是因为智库来源学者的关注领域与论文学者差异较大，论文学者关注科学问题较多，而智库学者关注经济、政治问题较多；也可能是因为地质学相关领域研究者的个人研究产出较高，但同领域研究者数量不及经济、政治等领域的学者数量，导致研究产出的总量低于经济、政治等领域。



图 4　学者研究领域词云

4.3.5　重要研究机构发现

对世界中国学专题数据库中的论文学者所属机构信息进行分析发现，在学术方面进行中国问题研究的热门机构，包括香港大学（UNIVERSITY OF HONG KONG）、加利福尼亚大学（UNIVERSITY OF CALIFORNIA）、香港城市大学（CITY UNIVERSITY OF HONG KONG）、香港理工大学（HONG KONG POLYTECHNIC UNIVERSITY）、新加坡国立大学（NATIONAL UNIVERSITY OF SINGAPORE）、南加利福尼亚大学（UNIVERSITY OF SOUTHERN CALIFORNIA）、澳大利亚国立大学（AUSTRALIAN NATIONAL UNIVERSITY）、香港科技大学（HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY）、莫纳什大学（MONASH UNIVERSITY）等院校、研究机构。关注中国问题相关学术研究的学者，可以持续关注这些热门的研究机构，跟进其研究进展。如图 5 所示：

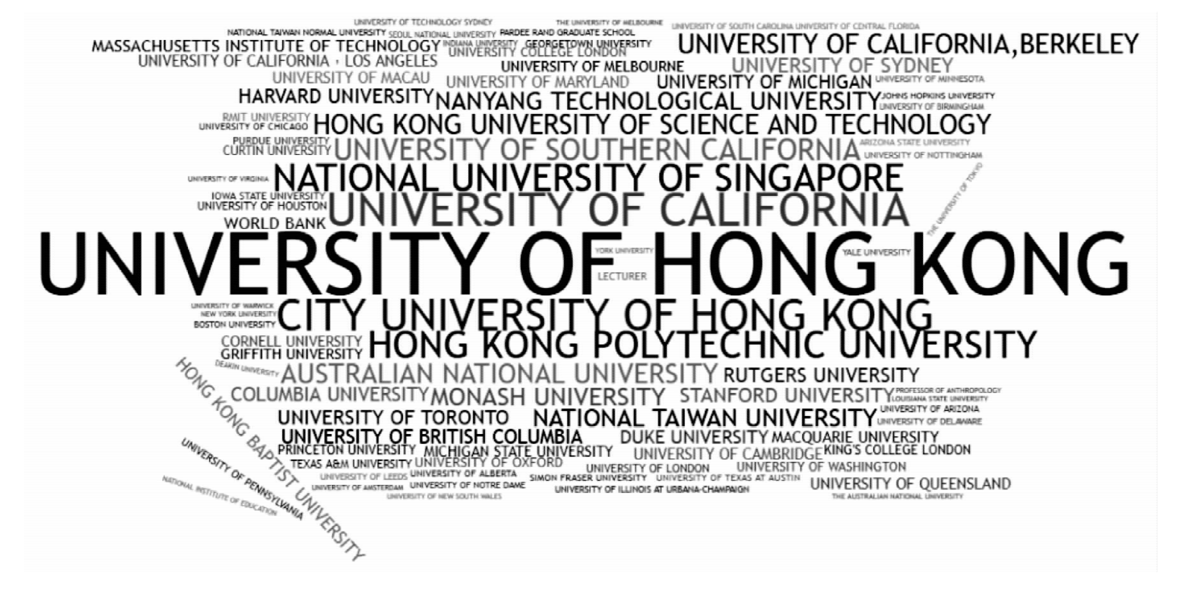


图 5　TOP100 学术研究机构

在世界中国学专题数据库中,依据学者数量对数据来源的 20 个智库进行排序及可视化展示,如图 6 所示。其中,Brookings(布鲁金斯学会)、RAND Corporation(兰德公司)、Chatham House(查塔姆研究所)位列前三。这 3 个机构的中国研究学者数占 20 个智库中中国研究学者总数的 39.58%,说明这 3 个智库对中国问题进行研究的学者较多,需要特别关注。

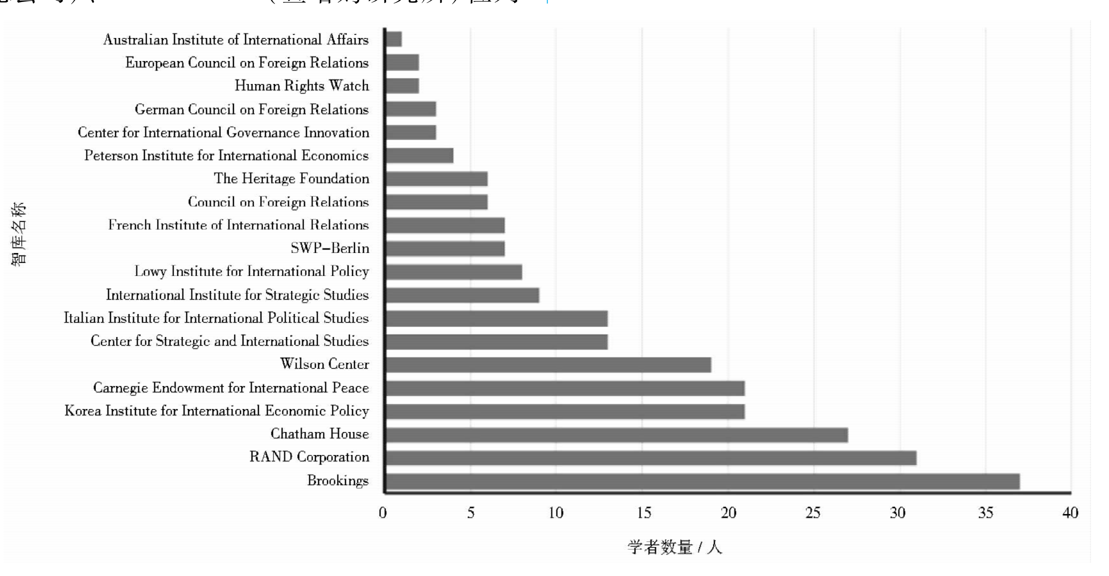


图 6　智库学者分布情况

5　结论与展望

建设世界中国学的专题数据库,在数据采集、加工整合及检索利用阶段相应解决当代世界中国学研究长期存在的难题,获取、整合和共享域外中国问题研究的知识资源,无论从学术还是现实层面都具有重大的意义。

华东师范大学世界中国学数据库,是通过构建世界中国学专题数据库应对当前世界中国学研究困境的一次尝试。在这次实践中,我们尝试将智库舆情上升到与学术研究同等的重视程度,同时关注传统"汉学"与当代中国问题的研究,而未生硬的将两类研究割裂开来。这是一次有意义的尝试,除了该数据库中已有的几类可视化分析,研究者还可以基于数据库中的数据开展更多视角、更多方法、更多维度的研究。需要注意的是,该数据库仍存在不足之处需要改进,包括:①数据采集时应用的检索词局限性大,这对于中国问题研究的广泛性和多样性而言,是非常片面的,导致目前呈现的数据体量与实际中国问题研究的体量有非常大的差距。这需要系统、深入地调研世界中国学的研究议题,设计出更为全面、体系、重点突出的检索词列表,以展示域外中国研究的全貌,避免重要议题的遗漏。②数据采集来源单一,如学术研究来源仅选取了谷歌学术,智库来源仅选择了 20 家智库机构,特别是舆情的重要来源——新闻媒体和社交网络,未包含在其中,这可能造成对重要研究机构和重要人物的识别与实际情况出现偏差。因此,需要对数据来源进行补充。可重点考虑受众广、影响大的域外主流媒体以及用户数量大、覆盖面广的社交网络作为舆情资源来源;在学术研究方面,要考虑非英语语种的学术资源,尤其补充日韩等邻国的学术资源。③多层次、多角度的数据关联、分析及可视化有限,如需获取更深层次的信息或结论,还需研究者依据数据自行进行挖掘和分析,直接服务决策的效果有限。因此,还需开发及应用更多的分析功能和可视化呈现。

从华东师范大学世界中国学数据库的建设和应用中可以看出,专题数据库的建设对该领域研究具有较大的促进作用,但在具体实践中仍存在一些问题。这些问题的发现为世界中国学专题数据库的未来持续优化和改进指明了方向,将促使其更加精准有效地服务于学术研究、服务于国家决策。

参考文献:

[1] 朱政惠. 史学理论与史学史研究的新思考——与海外中国学研究关系的讨论[J]. 国际汉学,2012(2):15-24.
[2] 梁怡. 习近平对外宣传重要论述与海外中国学研究[J]. 世界社会主义研究,2018(9):4-14,95.
[3] 上海社会科学院世界中国学研究所. 世界中国学理论前沿[M]. 上海:上海社会科学院出版社,2016:3.
[4] 朱政惠. 柯文教授的清史研究[J]. 江西师范大学学报(哲学社

会科学版),2004(6):25－31.
[5] 严绍璗. 对海外中国学研究的反思[J]. 探索与争鸣,2007(2):33－36.
[6] 吴原元. 百年来中国学人的域外汉学批评及其启示[J]. 史学理论研究,2019(4):112－124,159－160.
[7] 魏海生,唐洲雁,张树华. 面向新时代的海外当代中国研究[J]. 国外社会科学,2019(5):21－30.
[8] 张西平,石之瑜,罗曼诺夫,等. 70 年来的中国发展与海外中国研究[J]. 国外社会科学, 2019(6):61－73.
[9] 孙建军. 人文社科专题数据库建设关键技术与方法研究[J]. 信息资源管理学报, 2020,10(5):11.
[10] 施艳萍,李阳. 人文社科专题数据库关联数据模型的构建与应用研究[J]. 现代情报, 2019(12):19－27.
[11] 蒋婷,孙建军. 人文社科专题数据库深度语义化研究[J]. 信息资源管理学报, 2020,10(5):12－22.
[12] 毛平,剧晓红. 基于知识超网络的人文社科专题数据库数据资源聚合研究[J]. 信息资源管理学报,2020(5):38－47,54.
[13] 马鸿佳,李洁,沈涌. 数字资源聚合方法融合趋势研究[J]. 情报资料工作,2015(5):24－29.
[14] 董慧. 基于本体论和数字图书馆的信息检索[J]. 情报学报,2003(6):648－652.
[15] 2018 global go to think tank index report[EB/OL]. [2020－09－10]. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1017&context=think_tanks.

作者贡献说明:

邓璐芗:论文撰写;

许鑫:论文构思及审阅。

Construction and Application of the World China Studies Thematic Database
——A Case Study from East China Normal University

Deng Luxiang[1]　Xu Xin[1,2]
[1] Social Survey and Data Center, East China Normal University, Shanghai 200241
[2] Faculty of Economics and Management, East China Normal University, Shanghai 200062

Abstract: [Purpose/significance] Trying to build a thematic database of World China Studies to fix the problems including hard to get the knowledge of China from overseas, paying insufficient attention to public opinions and correlation of scholars in World China Studies. [Method/process] Introduce the design of the World China Studies thematic database in the aspects of data acquisition, data processing and integration, as well as data retrieval and utilization. And then did a case study of the World China Studies Thematic Database in East China Normal University. [Result/conclusion] Building the thematic database of World China Studies can greatly facilitate the locating of important scholars, important institutions, hot fields, as well as correlations among scholars. In addition, the database could also provide data resources for other in-depth and multidimensional research.

Keywords: world China studies　thematic database　think tank　public opinion